Foundations of Informatics: a Bridging Course

Part C: Context-Free Languages
Lesson 1: Introduction to Context-Free Grammars and Languages

Thomas Noll
Software Modeling and Verification Group
RWTH Aachen University

https://moves.rwth-aachen.de/people/noll/

Software Modeling and Verification Chair
RWTH AACHEN UNIVERSITY

Introductory Example I

Example

Syntax definition of programming languages by “Backus-Naur” rules

Here: simple arithmetic expressions

$$
\begin{array}{rcl}
\langle \mathit{Expression} \rangle & ::= & 0 \\
& \mid & 1 \\
& \mid & \langle \mathit{Expression} \rangle + \langle \mathit{Expression} \rangle \\
& \mid & \langle \mathit{Expression} \rangle * \langle \mathit{Expression} \rangle \\
& \mid & (\langle \mathit{Expression} \rangle)
\end{array}
$$

Meaning:

An expression is either 0 or 1, or it is of the form $u + v$, $u * v$, *or* (u) *where* u, v *are again expressions*

Introductory Example II

Example (continued)

Here we abbreviate $\langle \mathit{Expression} \rangle$ as E, and use "$\to$" instead of "$::=$". Thus:

$$E \;\to\; 0 \mid 1 \mid E + E \mid E * E \mid (E)$$

Introductory Example II

Example (continued)

Here we abbreviate $\langle \mathit{Expression} \rangle$ as E, and use "$\rightarrow$" instead of "$::=$". Thus:

$$E \rightarrow 0 \mid 1 \mid E + E \mid E * E \mid (E)$$

Now expressions can be generated by replacing nonterminal symbols according to rules, beginning with the start symbol E:

$$E \Rightarrow E * E$$

Introductory Example II

Example (continued)

Here we abbreviate $\langle \mathit{Expression} \rangle$ as E, and use "$\rightarrow$" instead of "$::=$". Thus:

$$E \rightarrow 0 \mid 1 \mid E + E \mid E * E \mid (E)$$

Now expressions can be generated by replacing nonterminal symbols according to rules, beginning with the start symbol E:

$$\begin{aligned} E &\Rightarrow E * E \\ &\Rightarrow (E) * E \end{aligned}$$

Introductory Example II

Example (continued)

Here we abbreviate $\langle \mathit{Expression} \rangle$ as E, and use "$\rightarrow$" instead of "$::=$". Thus:

$$E \rightarrow 0 \mid 1 \mid E + E \mid E * E \mid (E)$$

Now expressions can be generated by replacing nonterminal symbols according to rules, beginning with the start symbol E:

$$\begin{aligned} E &\Rightarrow E * E \\ &\Rightarrow (E) * E \\ &\Rightarrow (E) * 1 \end{aligned}$$

Introductory Example II

Example (continued)

Here we abbreviate $\langle \mathit{Expression} \rangle$ as E, and use "$\rightarrow$" instead of "$::=$". Thus:

$$E \rightarrow 0 \mid 1 \mid E + E \mid E * E \mid (E)$$

Now expressions can be generated by replacing nonterminal symbols according to rules, beginning with the start symbol E:

$$\begin{aligned} E &\Rightarrow E * E \\ &\Rightarrow (E) * E \\ &\Rightarrow (E) * 1 \\ &\Rightarrow (E + E) * 1 \end{aligned}$$

Example (continued)

Here we abbreviate $\langle \mathit{Expression} \rangle$ as E, and use "$\rightarrow$" instead of "$::=$". Thus:

$$E \;\rightarrow\; 0 \mid 1 \mid E + E \mid E * E \mid (E)$$

Now expressions can be generated by replacing nonterminal symbols according to rules, beginning with the start symbol E:

$$\begin{aligned} E &\Rightarrow E * E \\ &\Rightarrow (E) * E \\ &\Rightarrow (E) * 1 \\ &\Rightarrow (E + E) * 1 \\ &\Rightarrow (0 + E) * 1 \end{aligned}$$

Introductory Example II

Example (continued)

Here we abbreviate $\langle \mathit{Expression} \rangle$ as E, and use "$\rightarrow$" instead of "$::=$". Thus:

$$E \;\rightarrow\; 0 \mid 1 \mid E+E \mid E*E \mid (E)$$

Now expressions can be generated by replacing nonterminal symbols according to rules, beginning with the start symbol E:

$$\begin{aligned} E &\Rightarrow E*E \\ &\Rightarrow (E)*E \\ &\Rightarrow (E)*1 \\ &\Rightarrow (E+E)*1 \\ &\Rightarrow (0+E)*1 \\ &\Rightarrow (0+1)*1 \end{aligned}$$

Context-Free Grammars I

Definition

A context-free grammar (CFG) is a quadruple

$$G = \langle N, \Sigma, P, S \rangle$$

where

- N is a finite set of nonterminal symbols
- Σ is the (finite) alphabet of terminal symbols (disjoint from N)
- P is a finite set of production rules of the form $A \to \alpha$ where $A \in N$ and $\alpha \in (N \cup \Sigma)^*$
- $S \in N$ is a start symbol

Context-Free Grammars II

Example

For the above example, we have:

- $N = \{E\}$
- $\Sigma = \{0, 1, +, *, (,)\}$
- $P = \{E \to 0, E \to 1, E \to E + E, E \to E * E, E \to (E)\}$
- $S = E$

Context-Free Grammars II

Example

For the above example, we have:

- $N = \{E\}$
- $\Sigma = \{0, 1, +, *, (,)\}$
- $P = \{E \to 0, E \to 1, E \to E + E, E \to E * E, E \to (E)\}$
- $S = E$

Naming conventions:

- nonterminals start with uppercase letters
- terminals start with lowercase letters
- start symbol = symbol on LHS of first production

$\Rightarrow$ grammar completely defined by productions

Context-Free Languages I

Definition

Let $G = \langle N, \Sigma, P, S \rangle$ be a CFG.

- A sentence $\gamma \in (N \cup \Sigma)^*$ is directly derivable from $\beta \in (N \cup \Sigma)^*$ if there exist $\pi = A \to \alpha \in P$ and $\delta_1, \delta_2 \in (N \cup \Sigma)^*$ such that $\beta = \delta_1 A \delta_2$ and $\gamma = \delta_1 \alpha \delta_2$ (notation: $\beta \stackrel{\pi}{\Rightarrow} \gamma$ or just $\beta \Rightarrow \gamma$).

Context-Free Languages I

Definition

Let $G = \langle N, \Sigma, P, S\rangle$ be a CFG.

- A sentence $\gamma \in (N \cup \Sigma)^*$ is directly derivable from $\beta \in (N \cup \Sigma)^*$ if there exist $\pi = A \to \alpha \in P$ and $\delta_1, \delta_2 \in (N \cup \Sigma)^*$ such that $\beta = \delta_1 A \delta_2$ and $\gamma = \delta_1 \alpha \delta_2$ (notation: $\beta \stackrel{\pi}{\Rightarrow} \gamma$ or just $\beta \Rightarrow \gamma$).
- A derivation (of length $n \in \mathbb{N}$) of γ from β is a sequence of direct derivations of the form $\delta_0 \Rightarrow \delta_1 \Rightarrow \ldots \Rightarrow \delta_n$ where $\delta_0 = \beta$, $\delta_n = \gamma$, and $\delta_{i-1} \Rightarrow \delta_i$ for every $i \in \{1, \ldots, n\}$ (notation: $\beta \Rightarrow^* \gamma$).

Context-Free Languages I

Definition

Let $G = \langle N, \Sigma, P, S \rangle$ be a CFG.

- A sentence $\gamma \in (N \cup \Sigma)^*$ is directly derivable from $\beta \in (N \cup \Sigma)^*$ if there exist $\pi = A \to \alpha \in P$ and $\delta_1, \delta_2 \in (N \cup \Sigma)^*$ such that $\beta = \delta_1 A \delta_2$ and $\gamma = \delta_1 \alpha \delta_2$ (notation: $\beta \stackrel{\pi}{\Rightarrow} \gamma$ or just $\beta \Rightarrow \gamma$).
- A derivation (of length $n \in \mathbb{N}$) of γ from β is a sequence of direct derivations of the form $\delta_0 \Rightarrow \delta_1 \Rightarrow \ldots \Rightarrow \delta_n$ where $\delta_0 = \beta$, $\delta_n = \gamma$, and $\delta_{i-1} \Rightarrow \delta_i$ for every $i \in \{1, \ldots, n\}$ (notation: $\beta \Rightarrow^* \gamma$).
- A word $w \in \Sigma^*$ is called derivable in G if $S \Rightarrow^* w$.

Context-Free Languages I

Definition

Let $G = \langle N, \Sigma, P, S \rangle$ be a CFG.

- A sentence $\gamma \in (N \cup \Sigma)^*$ is directly derivable from $\beta \in (N \cup \Sigma)^*$ if there exist $\pi = A \to \alpha \in P$ and $\delta_1, \delta_2 \in (N \cup \Sigma)^*$ such that $\beta = \delta_1 A \delta_2$ and $\gamma = \delta_1 \alpha \delta_2$ (notation: $\beta \stackrel{\pi}{\Rightarrow} \gamma$ or just $\beta \Rightarrow \gamma$).
- A derivation (of length $n \in \mathbb{N}$) of γ from β is a sequence of direct derivations of the form $\delta_0 \Rightarrow \delta_1 \Rightarrow \ldots \Rightarrow \delta_n$ where $\delta_0 = \beta$, $\delta_n = \gamma$, and $\delta_{i-1} \Rightarrow \delta_i$ for every $i \in \{1, \ldots, n\}$ (notation: $\beta \Rightarrow^* \gamma$).
- A word $w \in \Sigma^*$ is called derivable in G if $S \Rightarrow^* w$.
- The language generated by G is $L(G) := \{w \in \Sigma^* \mid S \Rightarrow^* w\}$.

Context-Free Languages I

Definition

Let $G = \langle N, \Sigma, P, S \rangle$ be a CFG.

- A sentence $\gamma \in (N \cup \Sigma)^*$ is directly derivable from $\beta \in (N \cup \Sigma)^*$ if there exist $\pi = A \to \alpha \in P$ and $\delta_1, \delta_2 \in (N \cup \Sigma)^*$ such that $\beta = \delta_1 A \delta_2$ and $\gamma = \delta_1 \alpha \delta_2$ (notation: $\beta \stackrel{\pi}{\Rightarrow} \gamma$ or just $\beta \Rightarrow \gamma$).
- A derivation (of length $n \in \mathbb{N}$) of γ from β is a sequence of direct derivations of the form $\delta_0 \Rightarrow \delta_1 \Rightarrow \ldots \Rightarrow \delta_n$ where $\delta_0 = \beta$, $\delta_n = \gamma$, and $\delta_{i-1} \Rightarrow \delta_i$ for every $i \in \{1, \ldots, n\}$ (notation: $\beta \Rightarrow^* \gamma$).
- A word $w \in \Sigma^*$ is called derivable in G if $S \Rightarrow^* w$.
- The language generated by G is $L(G) := \{w \in \Sigma^* \mid S \Rightarrow^* w\}$.
- A language $L \subseteq \Sigma^*$ is called context-free (CFL) if it is generated by some CFG.

Context-Free Languages I

Definition

Let $G = \langle N, \Sigma, P, S \rangle$ be a CFG.

- A sentence $\gamma \in (N \cup \Sigma)^*$ is directly derivable from $\beta \in (N \cup \Sigma)^*$ if there exist $\pi = A \to \alpha \in P$ and $\delta_1, \delta_2 \in (N \cup \Sigma)^*$ such that $\beta = \delta_1 A \delta_2$ and $\gamma = \delta_1 \alpha \delta_2$ (notation: $\beta \overset{\pi}{\Rightarrow} \gamma$ or just $\beta \Rightarrow \gamma$).
- A derivation (of length $n \in \mathbb{N}$) of γ from β is a sequence of direct derivations of the form $\delta_0 \Rightarrow \delta_1 \Rightarrow \ldots \Rightarrow \delta_n$ where $\delta_0 = \beta$, $\delta_n = \gamma$, and $\delta_{i-1} \Rightarrow \delta_i$ for every $i \in \{1, \ldots, n\}$ (notation: $\beta \Rightarrow^* \gamma$).
- A word $w \in \Sigma^*$ is called derivable in G if $S \Rightarrow^* w$.
- The language generated by G is $L(G) := \{w \in \Sigma^* \mid S \Rightarrow^* w\}$.
- A language $L \subseteq \Sigma^*$ is called context-free (CFL) if it is generated by some CFG.
- Two grammars G_1, G_2 are equivalent if $L(G_1) = L(G_2)$.

Context-Free Languages II

Example

The language $\{a^n b^n \mid n \in \mathbb{N}\}$ is context-free. It is generated by the grammar $G = \langle N, \Sigma, P, S \rangle$ with

- $N = \{S\}$
- $\Sigma = \{a, b\}$
- $P = \{S \to aSb \mid \varepsilon\}$

(proof: generating $a^n b^n$ requires exactly n applications of the first and one concluding application of the second rule)

Context-Free Languages II

Example

The language $\{a^n b^n \mid n \in \mathbb{N}\}$ is context-free. It is generated by the grammar $G = \langle N, \Sigma, P, S \rangle$ with

- $N = \{S\}$
- $\Sigma = \{a, b\}$
- $P = \{S \to aSb \mid \varepsilon\}$

(proof: generating $a^n b^n$ requires exactly n applications of the first and one concluding application of the second rule)

Remark: illustration of derivations by derivation trees

- root labelled by start symbol
- leaves labelled by terminal symbols
- successors of node labelled according to right-hand side of production rule
- sequence of leaf symbols = generated word

Summary: Context-Free Grammars and Languages

Seen:

- Context-free grammars
- Derivations
- Context-free languages

Summary: Context-Free Grammars and Languages

Seen:

- Context-free grammars
- Derivations
- Context-free languages

Next:

- Relation between context-free and regular languages